JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 6
No. 1
April 2020

Analisis Cluster Terhadap Karakteristik Mahasiswa Jalur Prestasi FTI UKDW

R. Gunawan Santosa [#1], Antonius Rachmat Chrismanto[#2], Erick Kurniawan[*3]

[#]*Prodi Informatika, Fakultas Teknologi Informasi, Universitas Kristen Duta Wacana*
Jl. Dr. Wahidin Sudirohusodo No. 5-25 Yogyakarta
[1]gunawan@staff.ukdw.ac.id
[2]anton@ti.ukdw.ac.id

[*]*Prodi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Kristen Duta Wacana*
Jl. Dr. Wahidin Sudirohusodo No. 5-25 Yogyakarta
[3]erick@staff.ukdw.ac.id

***Abstrak*— Analisis Clustering merupakan analisis yang sangat bermanfaat untuk proses deskripsi dan eksplorasi sekumpulan data. Dengan tidak tahunya informasi terhadap karakteristik data mahasiswa jalur prestasi Universitas Kristen Duta Wacana (UKDW), maka akan sangat merugikan dalam pengambilan keputusan. Tujuan dari penelitian ini melihat karakteristik cluster masing-masing angkatan dari angkatan 2008 sampai dengan 2018. Dengan menggunakan analisis cluster, juga ingin diketahui kemiripan 10 angkatan mahasiswa jalur prestasi di FTI UKDW. Untuk melihat cluster tiap angkatan mahasiswa FTI UKDW digunakan metode K-Means Clustering. Sedangkan untuk menemukan kemiripan dari 10 angkatan mahasiswa FTI UKDW digunakan Hierarchial Clustering. Dari hasil penelitian didapat fakta bahwa sebagai berikut: dengan menggunakan K-Means Clustering untuk pengelompokkan menjadi dua cluster, maka diperoleh bahwa cluster yang mempunyai kecenderungan IP Semester 1 (IPS1) tinggi mempunyai karakteristik: status SMA swasta, lokasi SMA di Jawa, kategori SMA umum, level bahasa Inggris level 3, sedangkan cluster yang mempunyai kecenderungan IPS1 rendah mempunyai karakteristik: status SMA swasta, lokasi SMA luar Jawa, kategori SMA umum, level bahasa Inggris level 2. Apabila dilihat hasil pengelompokan tiap angkatan berdasarkan cluster yang terbentuk pada Dendrogram Hierarchical Clustering, maka Angkatan 2015 mempunyai kemiripan cluster yang paling berbeda dibandingkan dengan angkatan yang lainnya.**

***Kata kunci*— *K-Means Clustering*, *Hierarchical Clustering*, Dendrogram, Kemiripan.**

I. PENDAHULUAN

Perkuliahan di perguruan tinggi terdiri dari tiga tahap yaitu tahap awal perkuliahan, tahap pertengahan perkuliahan dan tahap akhir pekuliahan. Tahap awal perkuliahan sering disebut sebagai tahap adaptasi, tahap pertengahan merupakan tahap dimana mahasiswa sudah mulai terbiasa dengan cara belajar di perguruan tinggi dan pada tahap ini mahasiswa mulai menentukan bidang minat yang lebih spesifik melalui matakuliah yang telah diambilnya. Pada tahap akhir perkuliahan, mahasiswa mulai memikirkan hubungan antara pendidikan dan dunia kerja yang akan diambilnya.

Penelitian ini mempunyai cakupan yang berhubungan dengan masa awal perkuliahan di perguruan tinggi khususnya di UKDW, sebab hal tersebut merupakan masa peralihan dari sekolah menengah atas menuju pergurunan tinggi. Penelitian ini menggunakan metode analisis Clustering. Analisis Clustering merupakan analisis deskripsi dalam data mining dan selanjutnya kemudian akan dilakukan eksplorasi terhadap pola-pola yang ditemukan [1]. Data yang digunakan untuk analisis clustering berasal dari data mahasiswa angkatan 2008 sampai dengan angkatan 2018. Untuk membatasi permasalahan data tersebut hanya mencakup beberapa faktor, yaitu data kategori SMA, status SMA, lokasi SMA, Indeks Prestasi Semester 1 (IPS1) serta data level kemampuan bahasa Inggris untuk setiap mahasiswa.

Analisis clustering merupakan analisis yang penting dan banyak penelitian yang menggunakan analisis clustering. Analisis clustering dapat digunakan untuk beberapa tujuan tertentu, seperti misalnya pada beberapa penelitian Agustin W. & Erlin yang mengimplementasikan metode K-Means Clustering untuk membantu dalam memilih metode atau strategi promosi pada penerimaan mahasiswa baru. Strategi atau metode penerimaannya misalnya menggunakan media online, spanduk, map yang berisi brosur dan komunikasi lisan dengan adanya peran secara aktif dari mahasiswa dan para alumni. Hasil yang diperoleh adalah dapat dilihatnya strategi promosi mana yang lebih tepat dan terencana sebagai dukungan untuk bagian promosi dalam menggunakan strategi promosi yang efektif dan efisien [2].

Penelitian Waworuntu M. N. V. & Amin M. F. menerapkan metode K-Means Clustering dengan menggunakan sampel data sebanyak 440. Dari hasil perhitungan Davies Bouldin Index (DBI) diperoleh hasil

Submitted 28-10-2019; Revised 06-03-2020; Accepted 24-03-2020

bahwa penggunaan K-Means Clustering dengan hasil yang terbaik adalah dengan menggunakan nilai K=2. Penentuan beberapa nilai K diuji coba dan hasilnya adalah untuk K=2 cluster mempunyai nilai DBI = 0,243, untuk K=3 cluster mempunyai nilai DBI = 0,256, untuk K=4 cluster mempunyai nilai DBI = 0,275. Nilai yang yang terbaik dari beberapa nilai tersebut adalah untuk K=2 karena nilai tersebut mendekati 0. [3].

Metisen B. M. & Sari H. L. mengadakan penelitian dari data yang diolah dengan sampel data yang diambil di Swalayan Fadhilla Bengkulu. Penelitian ini menggunakan K-Means Clustering dan berhasil mengelompokan data barang yang terjual menjadi dua jenis kelompok data, yaitu data penjualan dengan frekuensi rendah dan data penjualan dengan frekuensi tinggi. Dengan adanya klasifikasi ini Swalayan Fadhila dapat mengetahui jenis barang yang laris terjual dan tidak laris terjual. Informasi dua kelompok barang ini sangat bermanfaat bagi perusahaan. Sehingga perusahaan dapat mempertimbangkan barang yang akan dibeli supaya tidak ada penumpukan barang yang ada di gudang [4].

Wanilah A.I. juga melakukan penambangan data melalui *clustering*. Metode yang diterapkan dalam data mining ini adalah K-means Clustering. Sedangkan dataset yang digunakan dalam clustering ini adalah data prestasi siswa. Prestasi siswa yang diteliti mempunyai beberapa atribut yaitu nama, ekstrakurikuler, nilai pengetahuan. Sedangkan nilai pengetahuan mencakup nilai sikap (perilaku), nilai keterampilan, dan presensi kehadiran siswa. Sampel datanya adalah 173 siswa dengan digunakan berbagai macam jenis perhitungan jarak, yaitu *Chebychep distance, Manhattan distance,* dan *Euclidean distance.* Persentase hasil akurasinya adalah 67%. Penelitian ini berhasil mengelompokkan siswa SMP Negeri I Sukahening berdasarkan prestasinya, yaitu cluster dengan prestasi "tinggi", "menengah" dan "cukup". [5].

Selain itu, Asroni & Adrian R. melakukan penelitian clustering dengan Weka Interface. Tujuan penelitian ini melakukan pengelompokan data siswa menjadi 4 cluster. Atribut yang digunakan pada dataset adalah NIM mahasiswa, nilai Algoritma dam Pemrograman 1, nilai Fisika Dasar, nilai Kalkulus 1dan IPK Dengan adanya cluster-cluster tersebut dapat memberikan rekomendasi siswa terbaik dari cluster yang sudah terbentuk. Cluster dengan anggota siswa yang terbaik dapat digunakan untuk memilih mahasiswa untuk mewakili kompetisi yang diadakan oleh Indonesia Security Incident Response Team on Internet Infrastructure (ID SIRTII) dari Kementrian Komunikasi dan Informatika RI [6].

Terakhir, Ezenkwu C.P., Ozuomba S. & Kalu C. mengadakan penelitian yang berhubungan dengan segmentasi pelanggan dengan metode K-Means Clustering. Tujuan dari penelitian ini adalah memberikan layanan pada pelanggan sesuai dengan target dan mengembangkan pemasaran yang sesuai untuk pelanggan. Atribut yang digunakan adalah rata-rata barang yang dibeli oleh pelanggan setiap bulannya dan rata-rata kunjungan pelanggan perbulan. Hasil dari penelitian ini diperoleh 4 cluster dengan akurasi 95%. Empat cluster tersebut adalah HBRV(High-Buyers-Regular-Visitor), HBIV(High-Buyers-Irregular-Visitor), LBRV(Low-Buyers-Regular-Visitor) dan LBIV(Low-Buyers-Irregular-Visitor) [7].

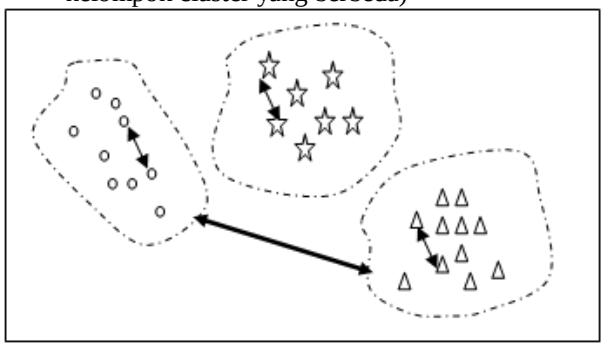
Adapun perbedaan antara penelitian yang dilakukan penulis dengan peneliti-peneliti sebelumnya adalah pada penelitian sebelumnya menggunakan dan menerapkan K-Means clustering pada beberapa dataset untuk menemukan pola pengelompokan yang ada, tetapi pada penelitian ini selain digunakan K-Means clustering pada setiap angkatan juga dilengkapi dengan Hiearchial Clustering untuk melihat pengelompokan pada tiap waktu yang berbeda.

Berdasarkan penelitian-penelitian yang telah dilakukan, maka tujuan penelitian yang akan dicapai ada dua. Yang pertama, melihat dan mendapatkan pola cluster berdasarkan centroid pada tiap angkatan dari angkatan 2008 sampai dengan 2018. Sedangkan yang kedua adalah melihat kemiripan tiap angkatan berdasarkan pembentukan centroidnya. Berdasarkan penelitian yang telah dibuat tersebut, ingin melakukan penerapan dua macam jenis metode clustering yang berbeda. Kedua metode tersebut tidak dapat dibandingkan mana yang lebih baik, tapi kedua metode tersebut akan saling melengkapi dalam menemukan pola suatu kumpulan data. Manfaat dari penelitian ini adalah melihat karakteristik cluster mahasiswa dengan IP semester yang "tinggi" sehingga bisa menjadi bahan pertimbangan bagi penerimaan mahasiswa baru. Penelitian ini merupakan kelanjutan dari penelitian [8] dan [9].

II. METODOLOGI PENELITIAN

Tujuan dalam clustering adalah untuk menemukan titik data yang secara alami (wajar) dikelompokkan secara bersamaan, dataset akan dipecah secara lengkap ke dalam beberapa cluster [8]. Dengan kata lain proses clustering mengandung beberapa hal, yaitu:

- Mengelompokkan data objek hanya berdasarkan informasi yang ditemukan dalam data yang menggambarkan objek–objek ini dan hubungan di dalamnya.
- Memaksimalkan kesamaan di dalam objek yang berada pada cluster yang sama
- Memaksimalkan perbedaan antara objek dalam kelompok cluster yang berbeda (atau dengan kata lain meminimalkan kesamaan antara objek dalam kelompok cluster yang berbeda)



Gambar. 1 Skema umum *clustering*

A. Clustering

Metode *clustering* merupakan metode yang sangat praktis .Metode ini sudah diterapkan di banyak bidang ilmu lain, seperti kecerdasan buatan dan pengenalan pola, *chemometric* , ekologi, ekonomi, *geosciences*, pemasaran, penelitian medis, ilmu politik, psikometri, dan banyak lagi. Saat ini telah muncul berbagai metode clustering dari yang sederhana sampai dengan yang rumit, seperti misalnya *Partitioning Around Medoids* (PAM), *Clustering Large Applications* (CLARA) , *Clustering* dengan *Fuzzy Analysis* (Funny), *Agglomerative Nesting* (AGNES) dan lain sebagainya. *Clustering* dikenal dengan berbagai nama, yaitu taksonomi numerik atau klasifikasi data otomatis [10].

Dalam analisis *cluster* dikenal adanya istilah *variate*. *Variate* dalam *cluster* adalah himpunan variabel yang mewakili karakteristik dan digunakan untuk membandingkan objek-objek dalam analisis *cluster*. Analisis *Cluster* adalah salah satu teknik dalam statistik multivariate yang tidak mempertimbangkan variasinya secara empiris tetapi analisis ini menggunakan variate seperti yang diinginkan peneliti.

Penekanan analisis *cluster* adalah pada perbandingan objek-objek berdasarkan variate yang digunakan peneliti, jadi bukan pada estimasi nilai variate itu. Selain itu, analisis cluster akan selalu membuat cluster terlepas dari "keberadaan sebenarnya dari setiap struktur dalam data". Kesesuaian adanya cluster-cluster sangat tergantung pada variabel atau faktor yang digunakan sebagai acuan untuk ukuran kesamaan atau kemiripan [10].

Analisis cluster biasanya dibagi menjadi dua metode besar, yaitu [11, 12]:

- Flat Cluster atau Partitional Clustering (PC) Partisi dalam metode ini tidak ada ketergantungan satu sama lain. Sebagai contohnya: K-means Clustering, Gaussian Mixture model.
- Hierarchical Clustering (HC) Partisi dapat divisualisasikan menggunakan struktur pohon (tree, dendrogram) serta tidak perlu jumlah cluster sebagai input. Sebagai contohnya: Agglomerative Clustering, Devisive Clustering.

Dalam tulisan ini akan digunakan K-Means Clustering untuk mewakili PC dan Agglomerative Hierarchical Clustering yang didasarkan pada dendrogram untuk mewakili HC.

B. K-Means Clustering

Ide *K-Means Clustering* pertama kali ditemukan oleh Lylod [10] kemudian disempurnakan oleh beberapa orang lagi. Metode K-means Clustering mempunyai bentuk algoritma sebagai berikut:

- **Input**: N Dataset $\{x_1, \ldots, x_N\}$ ($x_n \in R^D$); banyak cluster yaitu K
- **Initialize**: K centroid cluster yaitu $\mu_1, \ldots, \mu_K$.

Beberapa pilihan untuk initialization, yaitu:

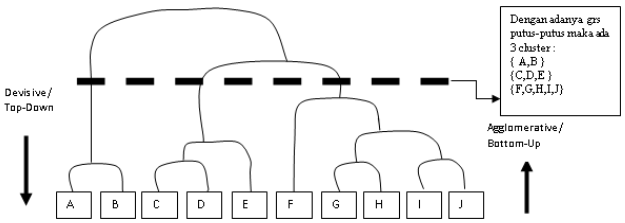
a) Memilih **secara random** sesuatu anggota dalam R^D
b) Memilih **sembarang K dari Dataset** sebagai centroid cluster

- **Iterate:**
  - **Assign** tiap-tiap x_n ke dalam centroid cluster yang terdekat
    $C_k = \left\{n : k = \arg\min_k \|x_n - \mu_k\|^2\right\}$, dimana ($C_k$ adalah himpunan yang jaraknya paling dekat dengan μ_k)
- **Recompute** centroid cluster yang baru untuk μ_k
  $\mu_k = \frac{1}{|C_k|}\sum_{n \epsilon C_k} x_n$
- **Repeat** bila belum konvergen

Kriteria yang mungkin untuk konvergen adalah centroid cluster tidak berubah lagi.

C. Hiearchial Clustering

Hierarchical clustering adalah pembuatan cluster yang memiliki urutan yang telah ditentukan dari atas ke bawah (bisa juga dari kiri ke kanan) . Sebagai contoh: semua file dan folder pada hard disk diatur dalam susunan hierarki. Ada dua jenis Hierarchical Clustering, yaitu Agglomerative Clustering dan Devisive Clustering [11]. Hierarchical Clustering biasanya menggunakan bantuan tree yang disebut dengan dendogram untuk membuat visualisasi clusternya [13].



Gambar. 2 *Hierachical clustering*

Dalam metode *Agglomerative Clustering* atau bottom-up, setiap observasi dipetakan menjadi clusternya sendiri. Lalu, hitung kesamaan (biasanya dengan menggunakan jarak) antara masing-masing cluster dan gabungkan dua cluster yang jaraknya paling dekat. Akhirnya, ulangi langkah 2 dan 3 sampai hanya ada satu cluster yang tersisa. Algoritma untuk Agglomerative dapat dilihat pada bagian dibawah ini:

Given
Himpunan X dari objek-objek $\{x_1, \ldots, x_N\}$ ($x_n \in R^D$)
Fungsi jarak dist(c_i, cj)
for i = 1 to N
    $c_i = \{x_i\}$
endfor
$C = \{c_1, \ldots, c_N\}$
I = N+1
C.size = I
while C. size > 1 **do**
    (c_{mini}, c_{minj}) = minimum dist(ci,cj) for all ci,cj in C

remove cmini and cminj from C
add {cmini, c_{minj}} to C
C.size = C. size - 1
I = I + 1
end while

Sebelum pengelompokan dilakukan, ditentukan terlebih dahulu matriks kedekatan yang berisi jarak antara setiap objek menggunakan fungsi jarak [13]. Jarak antara objek x_k dengan x_l adalah dengan persamaan 1.

$$dist(x_k, x_l) = \sqrt{\sum_{i=1}^{D}(x_{ki} - x_{li})^2} \quad (1)$$

Kemudian, matriks diperbarui untuk menampilkan jarak antara masing-masing cluster. Untuk mengukur jarak masing-masing cluster dapat digunakan Single Linkage, Complete Linkage atau Average Linkage [13]. Dalam hal ini hanya akan digunakan rumus Average Linkage.

Dalam Hierarchical Clustering hubungan jarak antara dua cluster didefinisikan sebagai rata-rata jarak antara setiap titik dalam satu cluster ke setiap titik dalam cluster lainnya. Misalnya, jarak antara cluster "r" dan "s" sama dengan rata-rata jarak setiap jarak yang menghubungkan titik-titik objek dari cluster r ke cluster s. Persamaan dari Average Linkage dapat dilihat pada Persamaan 2.

$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_i^r, x_j^s) \quad (2)$$

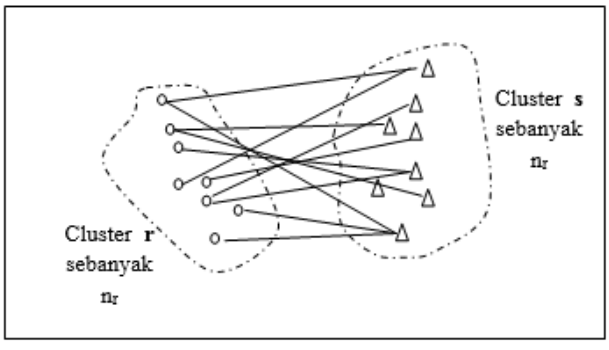
Catatan:

x_i^r adalah objek ke i dalam cluster "r"

x_j^s adalah objek ke j dalam cluster "s"

n_s dan n_r adalah masing-masing banyak anggota cluster "r" dan cluster "s".

Dapat dikatakan bahwa Average Linkage dalam *Hierarchical Clustering* merepresentasikan similaritas yang diukur dari rata-rata jarak dari semua objek dalam satu cluster ke semua objek di dalam cluster yang lain. Pendekatan ini cenderung akan dapat menggabungkan cluster dengan varians kecil [15].



Gambar. 3 Skema *average linkage*

Tahapan penelitian yang dilakukan pada penelitian ini adalah sebagai berikut :

1. Peneliti mengolah dataset dari data mahasiswa Fakultas Teknologi Informasi UKDW yang berasal dari PUSPINDIKA. Data mahasiswa yang digunakan untuk penelitian ini adalah data mahasiswa selama 11 tahun akademik, yaitu mahasiswa angkatan 2008 sampai dengan angkatan 2018.
2. Data dari PUSPINDIKA ini mencakup data Indeks Prestasi Semester 1 dari mahasiswa FTI, status SMA, kategori SMA dan lokasi SMA.
3. Kemudian peneliti juga akan mengambil data level kemampuan bahasa Inggris dari setiap siswa FTI UKDW dari PUSPINDIKA (dulu dari PPBA UKDW).
4. Mengintegrasikan data status SMA, lokasi SMA, kategori SMA dan Indeks Prestasi semester 1 dengan data level kemampuan bahasa Inggris untuk setiap siswa berdasarkan NIM mereka.
5. Membuat transformasi dari data tipe string menjadi tipe numerik
   Data kategori SMA (SMA menjadi "1" dan SMK menjadi "2") , status SMA (Negeri menjadi "1" dan Swasta menjadi "2") , lokasi SMA (Jawa menjadi "1" dan Luar Jawa menjadi "2") dan level bahasa Inggris (bhs Inggris level 1 menjadi "1" , bhs Inggris level 2 menjadi "2" , bhs Inggris level 3 menjadi "3" dan ESP menjadi "4") .
6. Memilah data mahasiswa FTI yang hanya melalui Jalur Prestasi (JP) saja, karena di UKDW proses penerimaan calon mahasiswa ada dua macam yaitu Jalur Prestasi (JP) dan Non Jalur Prestasi (NJP). Jalur NJP ini yang biasa disebut jalur reguler. Pada jalur prestasi tidak mengikuti Tes Kemampuan Akademik (TKA). Tes TKA ini mencakup tes kemampuan Numerik, Verbal, Spasial dan Analogi sedangkan pada Non Jalur Prestasi harus mengikuti empat tes kemampuan akademik tersebut.
7. Kemudian setelah proses *cleaning*, maka data dalam setiap angkatan tersebut dikelompokkan dengan metode K-Means Clustering dengan K=2 dan K=3. Pengelompokan dengan K-Means Clustering tersebut dengan menggunakan 5 variabel, sebagai berikut:
   a. x1 adalah status SMA
   b. x2 adalah lokasi SMA
   c. x3 adalah kategori SMA
   d. x4 adalah level bahasa Inggris
   e. x5 adalah IP semester 1

   Setelah proses K-Means Clustering akan didapat:
   - Centroid Cluster 21: centroid cluster 1, apabila kumpulan dataset dibagi menjadi dua cluster (ciri dari cluster ini adalah cluster dengan IP semester 1 yang tinggi)
   - Centroid Cluster 22: centroid cluster 2, apabila kumpulan dataset dibagi menjadi dua cluster (ciri dari cluster ini adalah cluster dengan IP semester 1 yang rendah)
   - Centroid Cluster 31: centroid cluster 1, apabila kumpulan dataset dibagi menjadi tiga

cluster (ciri dari cluster ini adalah cluster dengan IP semester 1 yang tinggi)

- Centroid Cluster 32: centroid cluster 2, apabila kumpulan dataset dibagi menjadi tiga cluster (ciri dari cluster ini adalah cluster dengan IP semester 1 yang sedang)
- Centroid Cluster 33: centroid cluster 3, apabila kumpulan dataset dibagi menjadi tiga cluster (ciri dari cluster ini adalah cluster dengan IP semester 1 yang rendah)

Pengelompokan antar angkatan dengan menggunakan metode Hierarchical Agglomerative Clustering dengan 30 variabel, seperti pada Tabel 1:

TABEL I
ATRIBUT UNTUK PROSES *AGLOMERATIVE CLUSTERING*

Var	Keterangan
x_1	status SMA centroid Cluster21
x_2	lokasi SMA centroid Cluster21
x_3	kategori SMA centroid Cluster21
x_4	level bhs Inggris centroid Cluster21
x_5	IP semester 1 centroid Cluster21
x_6	banyak anggota Cluster21
x_7	status SMA centroid Cluster22
x_8	lokasi SMA centroid Cluster22
x_9	kategori SMA centroid Cluster22
x_{10}	level bhs Inggris centroid Cluster22
x_{11}	IP semester 1 centroid Cluster22
x_{12}	banyak anggota Cluster22
x_{13}	status SMA centroid Cluster31
x_{14}	lokasi SMA centroid Cluster31
x_{15}	kategori SMA centroid Cluster31
x_{16}	level bhs Inggris centroid Cluster31
x_{17}	IP semester 1 centroid Cluster31
x_{18}	banyak anggota Cluster31
x_{19}	status SMA centroid Cluster32
x_{20}	lokasi SMA centroid Cluster32
x_{21}	kategori SMA centroid Cluster32
x_{22}	level bhs Inggris centroid Cluster32
x_{23}	IP semester 1 centroid Cluster32
x_{24}	banyak anggota Cluster32
x_{25}	status SMA centroid Cluster33
x_{26}	lokasi SMA centroid Cluster33
x_{27}	kategori SMA centroid Cluster33
x_{28}	level bhs Inggris centroid Cluster33
x_{29}	IP semester 1 centroid Cluster33
x_{30}	banyak anggota Cluster33

III. HASIL DAN PEMBAHASAN

A. Hasil

Dari pengolahan data angkatan 2008 sampai dengan 2018 dengan metode K-Means Clustering dan Hierarchical Agglomerative Clustering akan dibahas pada bagian-bagian selanjutnya. Hasil komputasi K-Means Clustering untuk K=2 dapat dilihat pada Tabel 2.

TABEL III
CENTROID DUA *CLUSTER K-MEANS CLUSTERING* DENGAN K=2 UNTUK MAHASISWA ANGKATAN 2008- 2018

Angkatan	Centroid	x_1	x_2	x_3	x_4	x_5	x_6
2008	Cluster_1	2	1	1	3	3,12	47
	Cluster_2	2	1	1	1	1,30	16
2009	Cluster_1	2	1	1	2	2,99	5
	Cluster_2	2	2	1	3	0,43	6
2010	Cluster_1	2	1	1	3	3,44	32
	Cluster_2	1	2	1	2	1,89	23
2011	Cluster_1	2	1	1	2	3,02	96
	Cluster_2	2	1	1	2	1,51	48
2012	Cluster_1	2	1	1	3	3,47	57
	Cluster_2	2	1	1	1	2,42	68
2013	Cluster_1	2	1	1	2	3,17	77
	Cluster_2	2	1	1	1	1,56	48
2014	Cluster_1	2	1	1	2	3,28	48
	Cluster_2	1	2	1	1	1,51	42
2015	Cluster_1	2	2	1	3	3,41	81
	Cluster_2	1	2	2	1	2,18	112
2016	Cluster_1	2	1	1	3	3,20	55
	Cluster_2	2	2	1	1	1,78	44
2017	Cluster_1	2	1	1	3	3,42	48
	Cluster_2	2	2	1	1	2,32	54
2018	Cluster_1	2	1	1	2	3,08	95
	Cluster_2	2	2	1	2	0,67	14
Rata-rata	Cluster_1	2,0	1,1	1,0	2,5	3,2	58,3
	Cluster_2	1,7	1,6	1,1	1,5	1,6	43,2

Hasil dari proses K-Means Clustering dengan K=3 dapat dilihat pada Tabel III.

TABEL IIIII
CENTROID TIGA *CLUSTER K-MEANS CLUSTERING* DENGAN K=3 UNTUK MAHASISWA ANGKATAN 2008- 2018

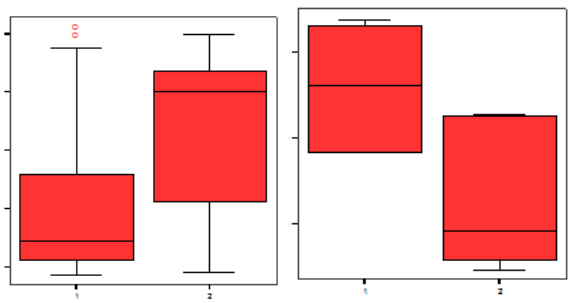
Angkatan	Centroid	x_1	x_2	x_3	x_4	x_5	x_6
2008	Cluster_1	1	1	1	3	3,28	22
	Cluster_2	2	1	1	2	2.82	33
	Cluster_3	1	2	1	2	0,29	8
2009	Cluster_1	2	1	1	2	3,27	4
	Cluster_2	2	2	1	2	0,70	5
	Cluster_3	1	1	1	4	0,46	2
2010	Cluster_1	2	1	1	3	3,43	20
	Cluster_2	1	1	1	2	2,95	27
	Cluster_3	2	2	1	2	0,68	8
2011	Cluster_1	2	1	1	3	3,12	38
	Cluster_2	2	1	1	2	2,71	88
	Cluster_3	1	2	1	2	0,29	18
2012	Cluster_1	2	1	1	3	3,54	29
	Cluster_2	2	1	1	2	2,98	85
	Cluster_3	2	1	1	2	0,64	11
2013	Cluster_1	2	1	1	3	3,27	30
	Cluster_2	2	1	1	2	2,93	63
	Cluster_3	2	2	1	1	1,13	32
2014	Cluster_1	2	1	1	3	3,46	19
	Cluster_2	2	1	1	2	2,94	44
	Cluster_3	1	2	1	1	0,93	27
2015	Cluster_1	2	2	1	3	3,53	36
	Cluster_2	1	2	1	1	3,01	111
	Cluster_3	1	2	2	1	1,27	46
2016	Cluster_1	2	1	1	3	3,15	27
	Cluster_2	2	1	1	2	2,86	53
	Cluster_3	1	2	1	1	0,91	19
2017	Cluster_1	2	1	1	3	3,31	29
	Cluster_2	2	1	1	2	3,12	60
	Cluster_3	2	1	1	1	0,49	13
2018	Cluster_1	2	1	1	3	3,24	35

Angkatan	Centroid	x_1	x_2	x_3	x_4	x_5	x_6
	Cluster_2	1	1	1	2	2,90	64
	Cluster_3	2	2	1	2	0,30	10
Rata-rata	Cluster_1	1,9	1,1	1,0	2,9	3,3	26,3
	Cluster_2	1,7	1,2	1,0	1,9	2,7	57,5
	Cluster_3	1,5	1,7	1,1	1,7	0,7	17,6

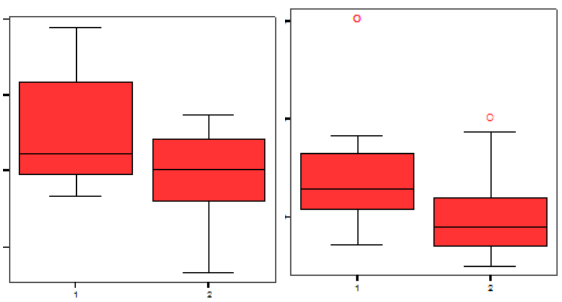
Untuk melihat ada tidaknya outlier pada tiap cluster dalam suatu angkatan dapat digunakan diagram BoxPlot. Diagram BoxPlot sering juga disebut dengan diagram Box dan Whiskers. BoxPlot adalah diagram yang menunjukkan pusat, penyebaran dan kemencengan sekumpulan data. Untuk menggambarkan diagram Box dan Whiskers dibutuhkan Kuartil 1, Median atau Kuartil 2, Kuartil 3, nilai data tertinggi dan nilai data terendah di antara Pagar Dalam Bawah (PDB) dan Pagar Dalam Atas (PDA).

Data pengamatan yang berada di luar PDB dan PDA disebut *outlier* atau pencilan. Pencilan ini dapat diklasifikasikan menjadi dua macam, yaitu outlier mild dan outlier ekstrim. Untuk melakukannya, kita mendefinisikan dua Pagar Luar, yaitu Pagar Luar Bawah (PLB) yang posisinya 3,0 IQR di bawah kuartil pertama dan Pagar Luar Atas (PLA) yang posisinya 3,0 IQR di atas kuartil ketiga. Jika data pengamatan berada di luar salah satu dari dua Pagar Dalam tetapi di dalam salah satu dari dua Pagar Luar, maka data itu akan disebut dengan *outlier* mild. Sedangkan data pengamatan yang berada di luar salah satu dari dua Pagar Luar disebut *outlier* ekstrim [14].

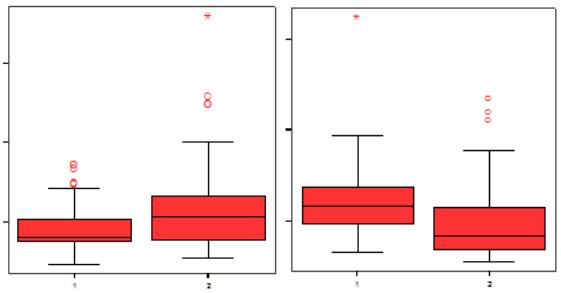
Untuk melihat ada tidaknya outlier dalam tiap cluster dapat digunakan diagram BoxPlot terhadap jarak suatu data dalam cluster yang diukur dari pusat clusternya. Gambar diagram BoxPlot untuk K=2 dapat dilihat pada Gambar 4.



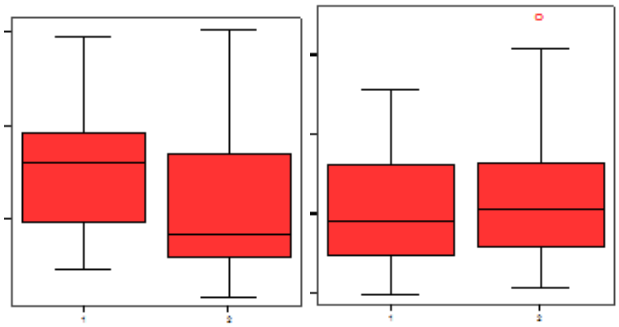
Gambar 4a. Angkatan 2008 dan 2009



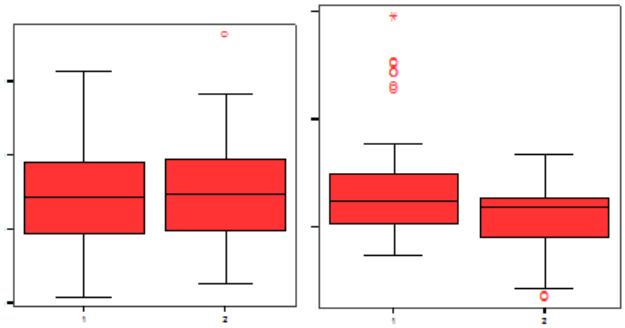
Gambar 4b. Angkatan 2010 dan 2011



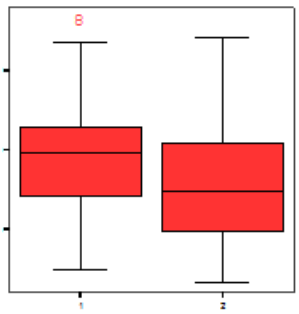
Gambar 4c. Angkatan 2012 dan 2013



Gambar 4d. Angkatan 2014 dan 2015



Gambar 4e. Angkatan 2016 dan 2017



Gambar 4f. Angkatan 2018

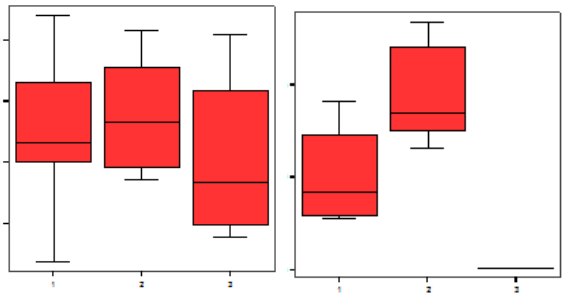
Gambar 4 (a) – (f). Diagram *BoxPlot* untuk *K-Means Clustering* dengan K=2 untuk Angkatan 2008 sampai dengan Angkatan 2018. Apabila seluruh gambaran visual itu diringkas, maka dapat dilihat pada tabel 4.

TABEL IVV
RINGKASAN ADANYA *OUTLIER* PADA *CLUSTER* DENGAN K=2 UNTUK MAHASISWA ANGKATAN 2008- 2018

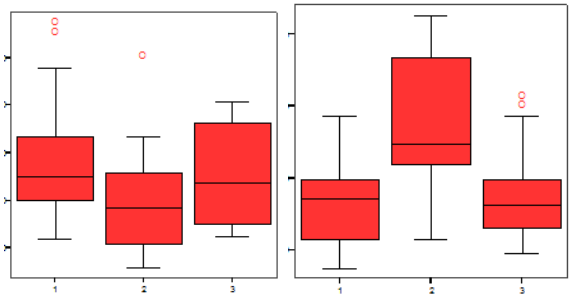
Angkatan	Karakteristik Outlier
2008	Cluster 1 mempunyai outlier mild Cluster 2 tidak mempunyai outlier
2009	Kedua cluster tidak mempunyai outlier
2010	Kedua cluster tidak mempunyai outlier
2011	Kedua cluster mempunyai outlier mild
2012	Kedua cluster mempunyai outlier mild dan bahkan cluster 2 mempunyai outlier ekstrim
2013	Cluster 1 mempunyai outlier ekstrim dan cluster 2 mempunyai outlier mild
2014	Kedua cluster tidak mempunyai outlier
2015	Cluster 2 mempunyai outlier mild
2016	Cluster 2 mempunyai outlier mild
2017	Cluster 1 mempunyai outlier mild dan bahkan outlier ekstrim Cluster 2 hanya mempunyai outlier mild
2018	Cluster 1 mempunyai outlier mild

Dari Tabel 4 terlihat bahwa untuk *K-Means Clustering* dengan K=2, pengelompokan pada angkatan 2009, 2010 dan 2014 yang terbaik karena masing-masing cluster dalam tiap angkatan tidak mempunyai outlier.

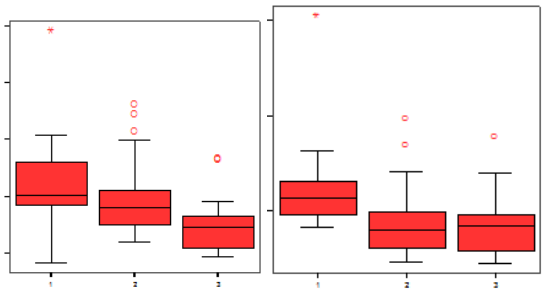
Gambar diagram *BoxPlot* untuk K=3 dapat dilihat pada Gambar 5. Apabila seluruh gambaran visual itu diringkas, maka dapat dilihat pada Tabel V.



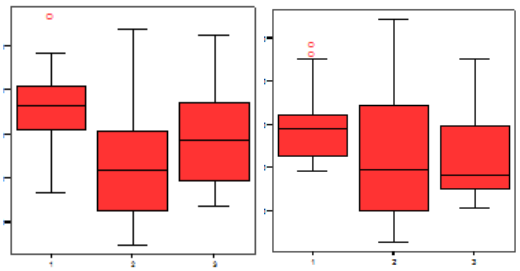
Gambar 5a. Angkatan 2008 dan 2009



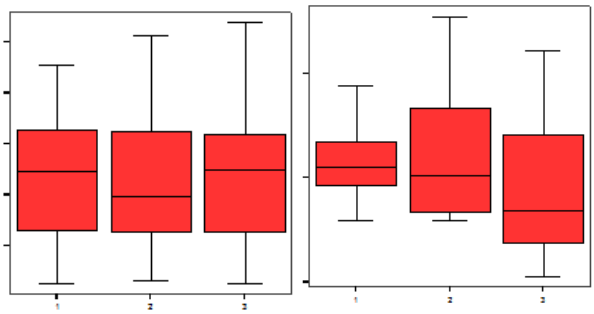
Gambar 5b. Angkatan 2010 dan 2011



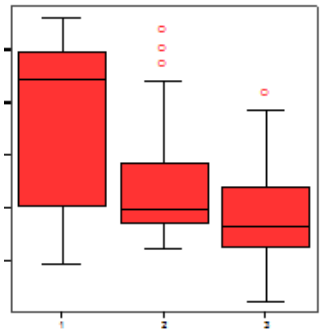
Gambar 5c. Angkatan 2012 dan 2013



Gambar 5d. Angkatan 2014 dan 2015



Gambar 5e. Angkatan 2016 dan 2017



Gambar 5f. Angkatan 2018

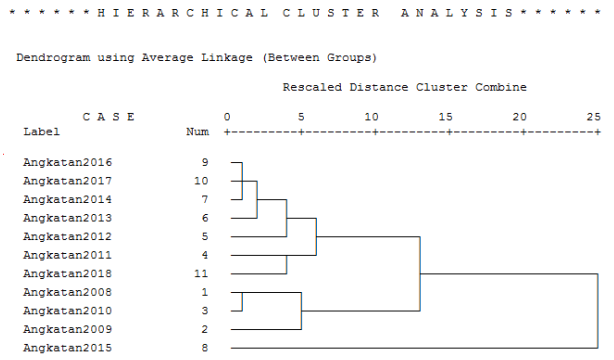
Gambar 5 (a)-(f). Diagram BoxPlot untuk K-Means Clustering dengan K=3 untuk Angkatan 2008 sampai dengan Angkatan 2018. Apabila seluruh gambaran visual itu diringkas, maka dapat dilihat pada tabel V.

TABEL V
RINGKASAN ADANYA *OUTLIER* PADA *CLUSTER* DENGAN K=3 UNTUK MAHASISWA ANGKATAN 2008- 2018

Angkatan	Karakteristik Outlier
2008	Ketiga cluster tidak mempunyai outlier
2009	Ketiga cluster tidak mempunyai outlier, cluster 3 tidak tergambar karena anggota clusternya hanya 2
2010	Cluster 1 dan cluster 2 mempunyai outlier mild Cluster 3 tidak mempunyai outlier
2011	Cluster 1 dan cluster 2 tidak mempunyai outlier Cluster 3 mempunyai outlier mild
2012	Cluster 1 mempunyai outlier ekstrim Cluster 2 dan cluster 3 mempunyai outlier mild
2013	Cluster 1 mempunyai outlier ekstrim Cluster 2 dan cluster 3 mempunyai outlier mild
2014	Cluster 1 mempunyai outlier mild Cluster 2 dan cluster 3 tidak mempunyai outlier
2015	Cluster 1 mempunyai outlier mild Cluster 2 dan cluster 3 tidak mempunyai outlier
2016	Ketiga cluster tidak mempunyai outlier
2017	Ketiga cluster tidak mempunyai outlier
2018	Cluster 1 tidak mempunyai outlier Cluster 2 dan cluster 3 mempunyai outlier mild

Dari tabel V terlihat bahwa untuk K-Means Clustering dengan K=3, pengelompokan pada angkatan 2008, 2009, 2016 dan 2017 yang terbaik karena masing-masing cluster dalam tiap angkatan tidak mempunyai outlier.

Hasil dari metode Hierarchical Agglomerative Clustering dengan menggunakan 30 variabel, dapat dilihat seperti pada pohon Diagram Dendogram pada Gambar 6.



Gambar. 6 Dendogram *average linkage*

B. Pembahasan

Metode K-means Clustering akan mengekstrak kumpulan data yang diwakili oleh centroid clusternya. Dengan menggunakan K-Means Clustering terlihat beberapa hal yang berhubungan centroid cluster-clusternya.

Dari Tabel II, untuk K-Means Clustering dengan K=2, didapat fakta bahwa:

- **Cluster 1** adalah cluster yang mempunyai ciri rata-rata IP Semester 1 yang tinggi (yaitu 3,00 sampai dengan 3,50), kecuali pada proses clustering Angkatan 2009 ada penyimpangan karena datanya sedikit, yaitu hanya 11 mahasiswa. Cluster 1 juga mempunyai kecenderungan x_1 (status SMA) = 2, x_2 (lokasi SMA) =1, x_3 (kategori SMA) = 1, x_4 (level bahasa Inggris) = 3.
- **Cluster 2** adalah cluster yang mempunyai ciri rata-rata IP Semester 1 yang rendah (yaitu 0,40 sampai dengan 2,50), kecuali pada proses clustering Angkatan 2009 ada penyimpangan karena datanya sedikit, yaitu hanya 11 mahasiswa. Cluster 2 juga mempunyai kecenderungan x_1 (status SMA) = 2, x_2 (lokasi SMA) =2, x_3 (kategori SMA) = 1, x_4 (level bahasa Inggris) = 2.

Dari Tabel III, untuk K-Means Clustering dengan K=3, didapat fakta bahwa:

- **Cluster 1** adalah cluster yang mempunyai ciri rata-rata IP Semester 1 yang tinggi (yaitu 3,10 sampai dengan 3,50). Cluster 1 juga mempunyai kecenderungan x_1 (status SMA) = 2, x_2 (lokasi SMA) =1, x_3 (kategori SMA) = 1, x_4 (level bahasa Inggris) = 3.
- **Cluster 2** adalah cluster yang mempunyai ciri rata-rata IP Semester 1 yang sedang (yaitu 2,70 sampai dengan 3,10), kecuali pada proses clustering Angkatan 2009 ada penyimpangan karena datanya sedikit, yaitu hanya 11 mahasiswa. Cluster 2 juga mempunyai kecenderungan x_1 (status SMA) = 2, x_2 (lokasi SMA) =1 , x_3 (kategori SMA) = 1, x_4 (level bahasa Inggris) = 2.
- **Cluster 3** adalah cluster yang mempunyai ciri rata-rata IP Semester 1 yang rendah (yaitu 0,20 sampai dengan 1,27). Cluster 2 juga mempunyai kecenderungan x_1 (status SMA) = 2, x_2 (lokasi SMA) =2, x_3 (kategori SMA) = 1, x_4 (level bahasa Inggris) = 2.

Untuk Hierarchical Clustering proses pembentukan clusternya dapat dilihat pada Gambar 4, sebagai berikut:

- Tahap 1: Case 9 dikelompokkan dengan Case 10, yaitu Angkatan 2016 dengan Angkatan 2017 (Case 9, Case10) = (Angkatan2016, Angkatan2017), sekarang disebut Case 9 = Angkatan (2016,2017)
- Tahap 2: Case 1 dikelompokkan dengan Case 3, yaitu Angkatan 2008 dengan Angkatan 2010 (Case 1, Case3) = (Angkatan2008,Angkatan2010) , sekarang disebut Case 1 = Angkatan(2008,2010)
- Tahap 3: Case 7 dikelompokkan dengan Case 9, yaitu Angkatan 2014 dengan Angkatan (2016,2017), sekarang disebut Case 7 = Angkatan (2014,2016,2017)
- Tahap 4: Case 6 dikelompokkan dengan Case 7, yaitu Angkatan 2013 dengan Angkatan (2014,2016,2017) disebut Case 6 = Angkatan (2013,2014,2016,2017)
- Tahap 5: Case 5 dikelompokkan dengan Case 6, yaitu Angkatan 2012 dengan Angkatan

(2013,2014,2016,2017) sekarang disebut Case 5 = Angkatan (2012,2013,2014,2016,2017)

- Tahap 6: Case 4 dikelompokkan dengan Case 11, yaitu Angkatan 2011 dengan Angkatan (2018), sekarang disebut Case 4 = Angkatan (2011,2018)
- Tahap 7: Case 1 dikelompokkan dengan Case 2, yaitu Angkatan (2008,2010) dengan Angkatan (2009) disebut Case 1 = Angkatan (2008,2009,2010)
- Tahap 8: Case 4 dikelompokkan dengan Case 5, yaitu Angkatan (2011,2018) dengan Angkatan (2012,2013,2014,2016,2017), sekarang disebut Case 4 = Angkatan (2011, 2012, 2013, 2014, 2016, 2017, 2018)
- Tahap 9: Case 1 dikelompokkan dengan Case 4, yaitu Angkatan (2008,2009,2010) dengan Angkatan (2011,2012,2013,2014,2016,2017,2018) , sekarang disebut Case 1 = Angkatan (2008,2009,2010,2011,2012,2013,2014,2016,2017, 2018)
- Tahap 10: Case 1 dikelompokkan dengan Case 8, yaitu Angkatan (2008,2009,2010,2011,2012,2013,2014,2016,2017, 2018) dengan Angkatan (2015) , sekarang disebut Case 1 = Angkatan(2008,2009,2010,2011,2012,2013,2014,2015,2016,2017,2018)

Dari Gambar 6, apabila setiap angkatan dibagi menjadi 2 cluster berdasarkan Average Linkage terlihat bahwa:

- **Cluster 1** terdiri dari {Angkatan 2008 sampai dengan Angkatan 2014 dan Angkatan 2016 sampai dengan Angkatan 2018}
- **Cluster 2** adalah {Angkatan 2015}

Demikian pula dari gambar Dendrogram Gambar 6, setiap angkatan dibagi menjadi 3 cluster, maka didapat:

- **Cluster 1** adalah {Angkatan 2008 sampai dengan Angkatan 2010}
- **Cluster 2** adalah {Angkatan 2011 sampai dengan Angkatan 2014 dan Angkatan 2016 sampai dengan Angkatan 2018}
- **Cluster 3** adalah {Angkatan 2015}

Secara garis besar perbedaan K-Means Clustering dan Hierarchical Clustering ada dua, yaitu:

- K-Means Clustering menghasilkan partisi tunggal sedangkan Hierarchical Clustering dapat memberikan partisi yang berbeda tergantung pada tingkat resolusi yang ingin dilihat.
- K-Means Clustering membutuhkan jumlah cluster yang harus ditentukan pada awalnya sedangkan Hierarchical Clustering tidak membutuhkan jumlah cluster yang ditentukan.

IV. KESIMPUILAN

Kesimpulan pertama yang diperoleh adalah apabila tiap angkatan dikelompokkan menjadi dua cluster, maka akan didapat bahwa cluster yang mempunyai kecenderungan IP Semester 1 tinggi mempunyai karakteristik: status SMA swasta, lokasi SMA di Jawa, kategori SMA umum, level bahasa Inggris 3, sedangkan cluster yang mempunyai kecenderungan IP Semester 1 rendah mempunyai karakteristik : status SMA swasta , lokasi SMA luar Jawa , kategori SMA umum , level bahasa Inggris 2.

Hasil berikutnya adalah K-Means Clustering dengan K=2 menghasilkan pengelompokan pada angkatan 2009, 2010 dan 2014 yang terbaik di antara angkatan yang lain, karena masing-masing cluster dalam tiga angkatan itu tidak mempunyai outlier, sedangkan K-Means Clustering dengan K=3 menghasilkan pengelompokan pada angkatan 2008, 2009, 2016 dan 2017 yang terbaik karena masing-masing cluster dalam empat angkatan itu tidak mempunyai outlier.

Terakhir, diperoleh bahwa apabila dilihat hasil pengelompokan dari tiap angkatan berdasarkan cluster yang terbentuk pada Dendrogram, maka Angkatan 2008 sampai dengan Angkatan 2014 dan Angkatan 2016 sampai dengan Angkatan 2018 berada dalam satu cluster dan mereka lebih mempunyai kemiripan dibandingkan dengan angkatan 2015 karena Angkatan 2015 mempunyai cluster dengan IP Semester 1 dengan kategori “sedang” yang paling banyak yaitu 111 orang.

UCAPAN TERIMA KASIH / ACKNOWLEDGMENT

Peneliti sangat berterima atas bantuan dana dan infrastruktur dari Fakultas Teknologi Informasi dan LPPM Universitas Kristen Duta Wacana Yogyakarta sehingga penelitian ini dapat berjalan dengan lancar.

REFERENSI

[1] J. Han and M. Kamber, Data Mining : Concepts and Techniques, Morgan Kaufmann Publishers, 2011.

[2] W. Agustin and E. , “Implementasi Metode K-Means Cluster Analysis untuk Memilih Strategi Promosi Penerimaan Mahasiswa Baru,” in *Seminar Nasional Ilmu Komputer (SNIK 2016)*, Semarang, 10 Oktober 2016.

[3] M. V. Waworuntu and M. F. Amin, “Penerapan Metode K-Means Pemetaan Calon Penerima JAMKESDA,” *Kumpulan Jurnal Ilmu Komputer (KLIK) Volume 05 No. 02*, pp. 190-200, September 2018.

[4] B. M. Metisen and H. L. Sari, “Analisis Clustering Menggunakan Metode K-Means dalam Pengelompokkan Penjualan Produk pada Swalayan Fadhila,” *Jurnal Media Infotama Vol. 11 No. 2*, pp. 110-118, September 2015.

[5] A. I. Warnilah, “Analisis Algoritma K-Means Clustering untuk Pemetaan Prestasi Siswa Studi Kasus SMP Negeri I Sukahening,” *Indonesian Journal on Computer and Information Technology Vol. 1 No. 1*, pp. 83-95, Mei 2016.

[6] Asroni and R. Adrian, “Penerapan Metode K-Means untuk Clustering Mahasiswa Berdasarkan Nilai Akademik dengan Weka Interface Studi Kasus pada Jurusan Teknik Informatika UMM Magelang,” *JURNAL ILMIAH SEMESTA TEKNIKA*, pp. 76-82, Mei 2015.

[7] C. P. Ezenkwu, S. Ozuomba and C. Kalu, “Application of K-Means Algorithm for Efficient Customer Segmentation : A Strategy for Targeted Customer Sevices,” *(IJARAI) International Journal of Advanced Research in Artificial Intelligence , Vol. 4 No. 10, 2015*, pp. 40-44, 2015.

[8] R. G. Santosa and A. R. Chrismanto, “Logistic Regression Model for Predicting First Semester Students GPA category Based on High School Academic Achievement,” *Researchersworld Journal*

of Arts, Science & Commerce Volume-VIII Issue-2(1) April 2017, pp. 58-66, 2017.

[9] R. G. Santosa and A. R. Chrismanto, "Perbandingan Akurasi Model Regresi Logistik untuk Prediksi Kategori IP Mahasiswa Jalur Prestasi dengan Non Jalur Prestasi," *jurnal Teknik dan ilmu Komputer Volume 07 No 25 Januari -Maret 2018*, pp. 107-121, 2018.

[10] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data : An Introduction to Cluster Analysis, New York: Wiley, 1990.

[11] A. C. Rencher, Methods of Multivariate Analysis Second edition, John Wiley & Sons, Inc. Publication, 2002.

[12] N. Jardine and R. Sibson, "The Construction of Hierarchic and Non-Hierarchic Classification," *Computer J. 11*, pp. 117-184, 1968.

[13] S. P. LLoyd, "Least Square Quantization in PCM," *IEEE Transactionon Information Theory 28(2)*, pp. 129-137, 1982.

[14] W. Kraus, D. Kraus and J. Lesinski, "A Simple Method of Construction and Rearrangement of Dendrogram," *COMPSTAT4, Physika Verlag, Vienna*, pp. 433-439, 1980.

[15] W. K. Hardle and L. Simar, Applied Multivariate Statistical Analysis 4th edition, Berlin: Springer-Verlag Berlin Heidelberg, 2015.

[16] J. F. Hair Jr, W. C. Black, B. J. Babin and R. E. Anderson, Multivariate Data Analysis Seventh Edition, Person New International Edition, 2014.

[17] P. S. Mann, Introductory Statistics Seventh Edition, John Wiley & Sons, Inc., 2010.